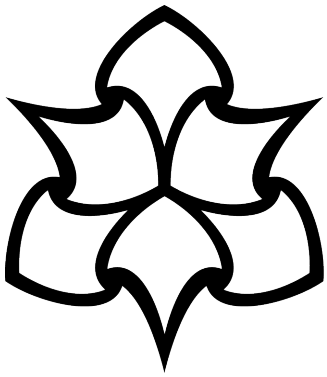

Studies in Higher Education

ISSN: 0307-5079 (Print) 1470-174X (Online) Journal homepage: https://www.tandfonline.com/loi/cshe20

A case for using both direct and indirect benchmarking to compare university performance metrics

W. E. Harris, A. M. Langan, N. Barrett, K. Jack, C. Wibberley & C. Hamshire

To cite this article: W. E. Harris, A. M. Langan, N. Barrett, K. Jack, C. Wibberley & C. Hamshire (2019) A case for using both direct and indirect benchmarking to compare university performance metrics, Studies in Higher Education, 44:12, 2281-2292, DOI: 10.1080/03075079.2018.1490893

To link to this article: https://doi.org/10.1080/03075079.2018.1490893

A case for using both direct and indirect benchmarking to compare university performance metrics

W. E. Harris, A. M. Langan, N. Barrett, K. Jack, C. Wibberley and C. Hamshire

Abstract

Benchmarking is used in higher education as a means to improve and compare performance. Comparative metric benchmarks may take two forms, based on direct standardization (DS) or indirect standardization (IS). DS can be used to measure variation in performance between institutions, controlling for intrinsic differences at each institution (e.g. controlling for differences in student typologies). IS can be used to measure variation in performance between institutions, compared to average performance overall. Typically, IS has been used to moderate educational output metrics, such as student qualification and satisfaction. We contrast the two approaches with an example dataset for three years of nursing student completion rates from nine institutions. Profiles of benchmarks and actual performance indicated that both approaches provide valuable and different perspectives to comparisons of institutional performance. We discuss the potential merits to stakeholders of each approach and conclude that decision-making can be best informed using both benchmark methods.

Background

Performance metrics are an established approach for evaluating and comparing institutional performance in higher education (HE), at local, national and international levels (Hazelkorn 2015; Chinta, Kebritchi, and Elias 2016). This has become increasingly important both from an institutional perspective (e.g. by informing quality and performance management) and for other HE stakeholders, such as funding bodies and prospective students (Hazelkorn 2007). Public classifications of educationally derived output metrics, for example in the form of league tables, have high impact on institutional reputation on a global scale (Dill and Soo 2005). Despite the widespread use of output metrics, there is a lack of consensus for standard calculation methods and many competing approaches have been described (e.g. Tam 2001; Chinta, Kebritchi, and Elias 2016). Direct quantitative comparison of institutions is hampered by failing to account for the impact of variation in student typologies on performance measures. This exacerbates concerns over the acceptance of simple, ranked approaches as measures of 'academic quality', and the associated pressure to increase absolute and rank measures per se, as opposed to increasing quality (Bowden 2000).

An additional layer of HE quality and performance measures is the use of metric benchmarking. Metric benchmarking has an advantage over dissemination of simple ranking of absolute metrics in that it provides a weighted measure of performance that is directly comparable to performance in other institutions. Benchmarking is an established practice in business performance management

(Kumar and Chandra 2001) and has resulted in an array of benchmarking definitions and approaches (Zairi 1998; Kyrö 2003). There is evidence for the usefulness of metric benchmarking in HE (Agasisti and Bonomi 2014); however, potential benefits are subject to the challenges of effectively implementing benchmarking, or other performance metrics, as a quality enhancement tool (Hazelkorn 2015).

Jackson (2001) outlines the history of benchmarking in HE, beginning in North America in the 1980s, where it was used as a management tool for non-academic services in HE. The practice spread and broadened in scope to become a prominent way to manage academic quality and focussed on HE accountability and international competitiveness. Benchmarking has proliferated across the sector and transformed HE management globally (Hazelkorn 2015, 42). Many approaches to benchmarking have been documented and the literature provides a history of nuanced definitions surrounding different applications (e.g. Jackson and Lund 2000). A key challenge of benchmarking is to devise metrics to directly compare 'processes with outcomes', usually amongst institutional entities with comparable goals. However, there remains ongoing debate about the processes, comparative metrics and data that should be used (e.g. Tee 2016). This is possibly because there are many ways to categorize benchmarking approaches, and JISC (2012) provides a distinction between 'metric' and 'process' benchmarking in education and research. Metric benchmarking provides information to identify significant performance gaps, whereas process benchmarking uses metric benchmarks as a basis for understanding performance gaps through examination and comparison of processes. Ultimately, the method of calculation of the underpinning metric benchmark is critical to inform subsequent processes and decision-making.

Measurement of performance or efficiency in HE often relies on data relating to the outputs of degree courses, such as qualification rates, student employment outcomes and levels of student satisfaction. While these are metrics that students and HE academics and managers are interested in, some metrics (and other factors associated with them that could aid interpretation) can be difficult to obtain reliably and consistently for comparison between institutions (Williams and de Rassenfosse 2016). The biggest advantage of metric benchmarks over absolute metrics or simple ranking schemes is that they can account for variation in the student population, including social and economic factors, when comparing institutional performance. This overcomes the significant challenge of accounting for diversity when considering institutional performance, not only in HE but also in other education and employment sectors (Pitts 2005).

Equitability in HE benchmarking systems that compare institutions requires adjusted outcomes to account for inherent differences of the institutions, which may include heterogeneity of students, staff and programmes of the study (e.g. see HESA 2011). Any adjustments to absolute values should be made transparent by practitioners or managers, and context should be provided for interpretation. However, many stakeholders may be unaware of the wider context of benchmarks or the different approaches available for their calculation. There are two broad approaches to metric benchmarking, called 'direct' and 'indirect' standardization (Draper and Gittoes 2004; see below), and while both approaches can use the same institutional data, they are different in implementation and interpretation.

The basis of benchmarking in HE relies on the identification of characteristics for comparison between institutions. This could be represented by a variety of features, such as performance within and between programme subjects, or, as we highlight here, social or demographic features of the student cohort represented at different institutions that can be associated with performance measures, such as satisfaction or qualification. An initial challenge is the identification of data representing constituent typologies (of students) represented in institutions (Asif and Raouf 2013). This information must be readily available and comparable for all institutions to be benchmarked, which is true of the basic demographic information used across international HE institutions. Direct standardization (DS) benchmarking weighs the overall performance of student groupings across institutions and estimates the performance anticipated were this aggregate student population to attend each individual participating institution (Figure 1(a)). The expected DS benchmark

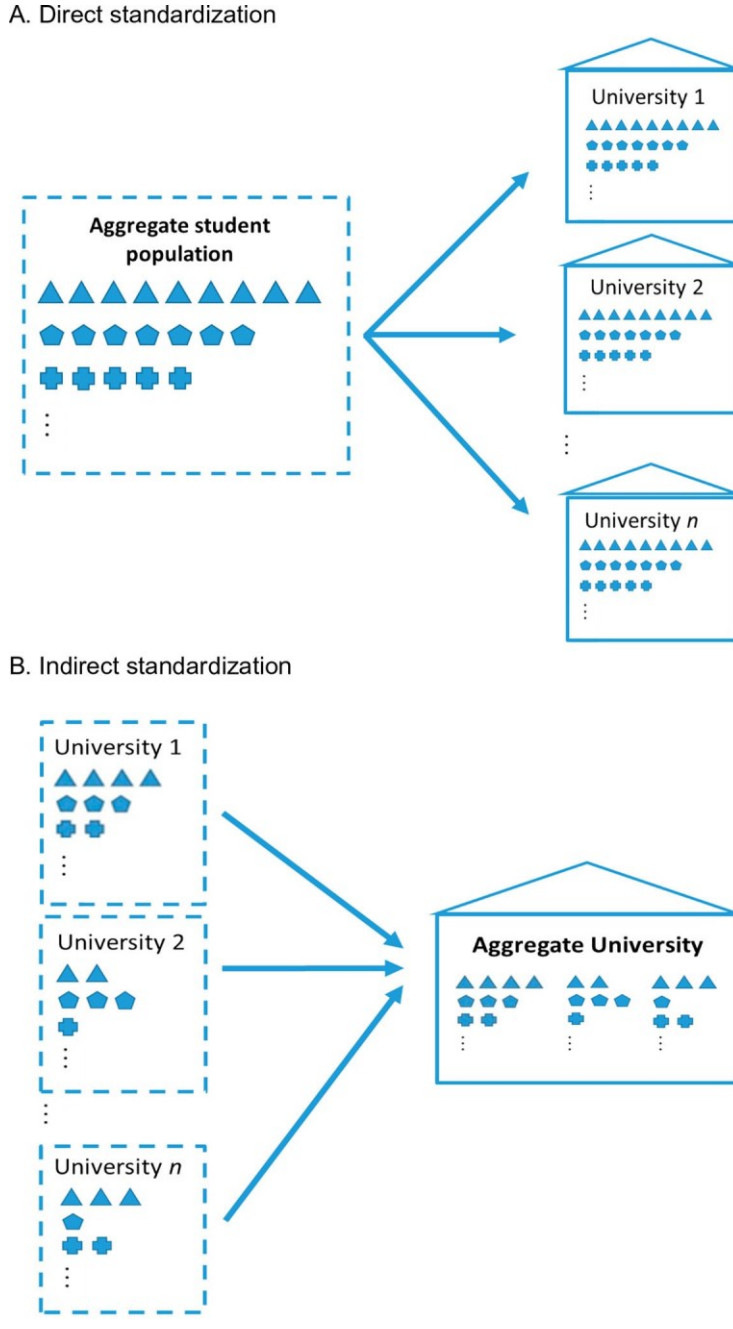

Figure 1. Conceptual explanation of direct and indirect benchmarking. Each symbol represents a different student demographic segment, with the number of each individual symbol representing the proportion of the student segment in that particular population. For example, if the variable's age and gender are considered, the triangles might represent the subset of 'young' AND 'female' students, the pentagonal symbols might represent 'mature' AND 'female', etc., with the other symbols representing different demographic groupings, respectively. A typical benchmark analysis would include many such groupings, the number of which would depend on the number of grouping variables and the number of combinations of levels explicitly considered in the benchmark. For simplicity, here we show only three different groupings by way of example. (a) Direct standardization approximates the expected performance of the aggregate student demographic segments at individual universities. (b) Indirect standardization approximates the performance of separate cohorts of students based on expected average performance.

performance can then be compared to the actual performance of the cohort of students, and their varying demographic representation, at the individual institutions. Here, the benchmark per se is the difference between the actual performance of the aggregate student cohort at that institution relative to the performance of the actual local student cohort.

Indirect standardization (IS) benchmarking, in contrast to DS, has been used widely in HE (Draper and Gittoes 2004), for example to compare student evaluation quality across institutions in the UK National Student Survey (Fielding, Dunleavy, and Langan 2010). Here, weighted averages of output metrics are calculated to account for the different compositions of student typologies across different student groupings, such as subject or degree programme. This can be used to evaluate achievement of students from different backgrounds (e.g. socio-economic, ethnicity, age), which may be heterogeneously represented across institutions, and may reveal different levels of achievement on a national level exhibited by different student populations. The general approach for IS is to weigh the aggregate performance of individual student groupings based on identified demographic parameters on a 'per institution' basis, relative to the aggregate performance of the same groupings across all institutions. Conceptually, this can be described as an 'aggregate university' (Figure 1(b)). For example, if national data are used this could be conceptualized as comparing individual aggregations of students at their respective institutions to the performance of the aggregate student population at a 'national university'. The goal is that benchmarking, in this sense, 'levels the playing field' to allow more meaningful comparisons between institutions. The expected IS performance may then be compared to the actual local performance measures to explore whether courses, or institutions, over- or underperform based on their student composition.

While there is some literature describing these different benchmarking approaches in HE (Draper and Gittoes 2004), there is a paucity of literature surrounding the application and interpretation of the two approaches. The education and health sectors have been particularly subjected to an increase in the use of performance measures as an indicator of quality (Northcott and Llewellyn 2005; Shober 2013), and previous work has demonstrated and highlighted the importance of factor selection and differences in performances of indirectly standardized benchmarks of regional nursing courses (Langan et al. 2016). We suggest that there is an imperative to clarify the comparative benefits and interpretation of the two benchmarking approaches, both for stakeholders and for practitioners. In this study, we use three years of student completion data from nursing courses at nine universities in the UK. Our aim was to compare performance of individual institutions using both DS and IS benchmarking techniques and provide guidance on the interpretation of the outcomes.

Methods

Dataset

The dataset used encompassed over 36,000 students in allied healthcare training programmes beginning their degree in the academic years 2008−2011 and who would have become qualified in the field of study by 2014 at the earliest. Nine participating institutions located in the north of England were included, and the data were anonymized according to guidelines imposed by Health Education England (formerly Health Education North West). The dependent variable was binary, indicating whether or not students achieved a qualification in their field of training subsequent to graduation. A number of descriptor variables were available to analyze that contained information associated with individual students, including factors such as age, gender, ethnicity, disability status and whether or not the student had suspended their studies. The anonymized baseline dataset contained information associated with individual students routinely recorded as part of the Professional Education Training Database. Further socio-economic information was associated with the student home postcode, such as youth and adult participation in further and HE arising from the Higher Education Funding Council for England dataset (i.e. HEFCE POLAR data; see Harrison and Hatt 2010).

Even for large datasets, benchmarking is constrained to use a relatively small number of explanatory factors to create different data groupings for the analysis (Hall and Holmes 2003). The factors selected for inclusion are important both for the information content of any benchmarking model and also for interpretation of results (Draper and Gittoes 2004; Langan et al. 2016). Here, one or more variables can be thought of as representing different demographic 'segments' of the student population. The baseline dataset contained 52 explanatory variables, from which five were chosen to represent student demographic segments: age, gender, ethnicity, disability status and youth participation in HE associated with home postcode. The selection criteria for these variables for this dataset have previously been described (Langan et al. 2016).

A machine learning method (Random Forest analysis) was used to select the most important variables to explain variation in the dependent variable, in this case student qualification (Breiman 2001; Hapfelmeier and Ulm, 2013). Random forest is a popular and efficient algorithm for large datasets suited to classification and regression analyses (Genuer, Poggi, and Tuleau 2010). The criterion used for ranking importance was a weighted average of those with the highest standardized mean for node purity (how much qualification classification changes when a target variable is excluded from analysis), and the percentage of variation explained in the dependent variable. Briefly, this analysis reduced the dataset by identifying a subset of the most useful variables for benchmarking based on objective criteria. This method is also considered robust to the inclusion of heterogeneous data, such as student demographic information (Verikas, Gelzinis, and Bacauskiene 2011). This resulted in five variables being identified as having large explanatory power for our dependent variable, student qualification. The same demographic variables were used for both benchmarking models in order to facilitate direct comparison of results. All machine learning analysis for factor selection was performed in the statistical package 'R' (R Core Team 2017) using the approach described by Liaw and Wiener (2002).

Benchmarking

DS and IS benchmarks were calculated according to the general approach described by Draper and Gittoes (2004). For both benchmarking approaches, this entails the calculation of a weighted mean of the student qualification rate within and between the student demographic segments described above. For the DS benchmarking approach, we calculated the mean qualification rate for demographic segments in aggregate across participating institutions. We then weighted the segment-specific mean qualification rate by the representation of those demographic segments for each individual institution, based on the actual performance of those demographic segments at each respective institution. For the IS benchmarking approach, we calculated the performance of the student demographic segments averaged across the individual institutions. For both these methods, the standardized expected performance was then compared to the actual qualification performance for each anonymized institution.

Factor number, factor order and the evenness of distribution of students across the different demographic segments all can have a strong effect on the outcome of benchmarking, and thus were explicitly considered. Our approach for factor order and number was to perform sensitivity analysis on different possible combinations of factor order and factor number to include in our models. The details of the sensitivity analysis have been described previously for the IS benchmarks (see Langan et al. 2016), but are briefly reported here for context. All possible combinations for order of inclusion of the five factors were considered in competing benchmark models. In comparing benchmark outcomes based on the different factor orders for an IS model, no qualitative difference in benchmarking outcomes relative to actual performance was found. However, there was a small effect of factor order on the variation in benchmarking outcomes (i.e. some institutions exhibited relatively smaller or larger variance in benchmark estimates; see Langan et al. 2016).

We manipulated the number of factors included in competing benchmarking models. This is an important consideration, especially when data across combinations of factors do not contain a

similar number of observations. We calculated competing benchmarks for all possible numbers of factors (i.e. 5, 4, 3, 2 or 1). Since factor order could interact with factor number, within each number of factors tested we compared all possible orders of factors as well. In comparing benchmarking outcomes based on factor number, no qualitative difference was found. However, as with the effect of factor order for five factors, there was a small effect of factor number on variation in benchmarking outcomes. The effects of factor order and factor number were arbitrary with respect to the qualitative outcome and all five candidate variables were included and the factor order we used for further analysis was based on age, gender, ethnicity, registered disability and youth participation rate in HE based on home post code. To facilitate direct comparison with IS benchmarks, the DS benchmarks we report here were calculated using exactly the same factor number and order as the IS benchmarks previously reported (Langan et al. 2016).

The deviation between benchmarks and actual performance for both DS and IS methods were calculated and the overall mean and standard deviation were used for comparative purposes. We compared our results to the criterion of a 0.03 difference between actual and benchmark performance to indicate significance as 'a rule of thumb'. We adopted this threshold as it has been used as a standard to indicate significant differences in previous HE benchmark studies (HEFCE 1999; Draper and Gittoes 2004).

Findings

DS benchmark results are shown in Figure 2(a). Overall, mean DS results differed from actual student qualification results by 1.44% (standard deviation ±1.81). In our sample of nine universities, three institutions had DS benchmark values higher than actual performance, five had DS benchmark values lower than actual performance, and one exhibited no difference between the DS benchmark and actual performance. Only one institution differed by more than the rule of thumb threshold for significance of 0.03; however, three universities had DS benchmark performance that varied by greater than one standard deviation from actual performance ('D', Figure 2(a)). IS results are shown in Figure 2(b). Mean IS differed from actual performance by 5.78% (standard deviation ±8.10). In our sample, four universities had higher IS benchmark performance than actual performance, with three greater than one standard deviation. Three universities had lower IS than actual performance (two greater than one standard deviation lower), and two universities had benchmark performance that did not differ from actual performance. Here, five universities had indirect benchmark performance that differed significantly according to our criterion of 0.03.

We found inconsistency between benchmark methods in the deviations in performance for specific institutions. Only one institution ('D', Figure 2(a,b)) had large deviations in both direct (greater than one standard deviation) and indirect (greater than 3%) benchmarks. Six institutions ('A', 'B', 'C', 'E', 'F', 'H', and 'I') have large deviations from actual performance in at least one benchmark measure, with the remaining institutions ('E' and 'G') showing only small deviation from actual performance. There was no correlation between IS and DS benchmark differences (Spearman's $\rho = -0.12$, $df = 8$, $P = .76$). There also was no correlation between relative actual performance and the DS benchmark (Spearman's $\rho = -0.23$, $df = 8$, $P = .55$); however, we found a significant negative correlation between relative actual performance and the IS benchmark (Spearman's $\rho = -0.83$, $df = 8$, $P = .0058$).

Interpretation

This section is intended to provide an explicit example comparing the two benchmarking methods based on the relative performance of specific institutions. We found contrasting outcomes for institutions when benchmarked by DS and IS methods, compared to actual performance. There was no correlation between the two benchmark results, suggesting differences in the information each benchmark provides. While IS benchmarks and actual performance were strongly correlated, DS

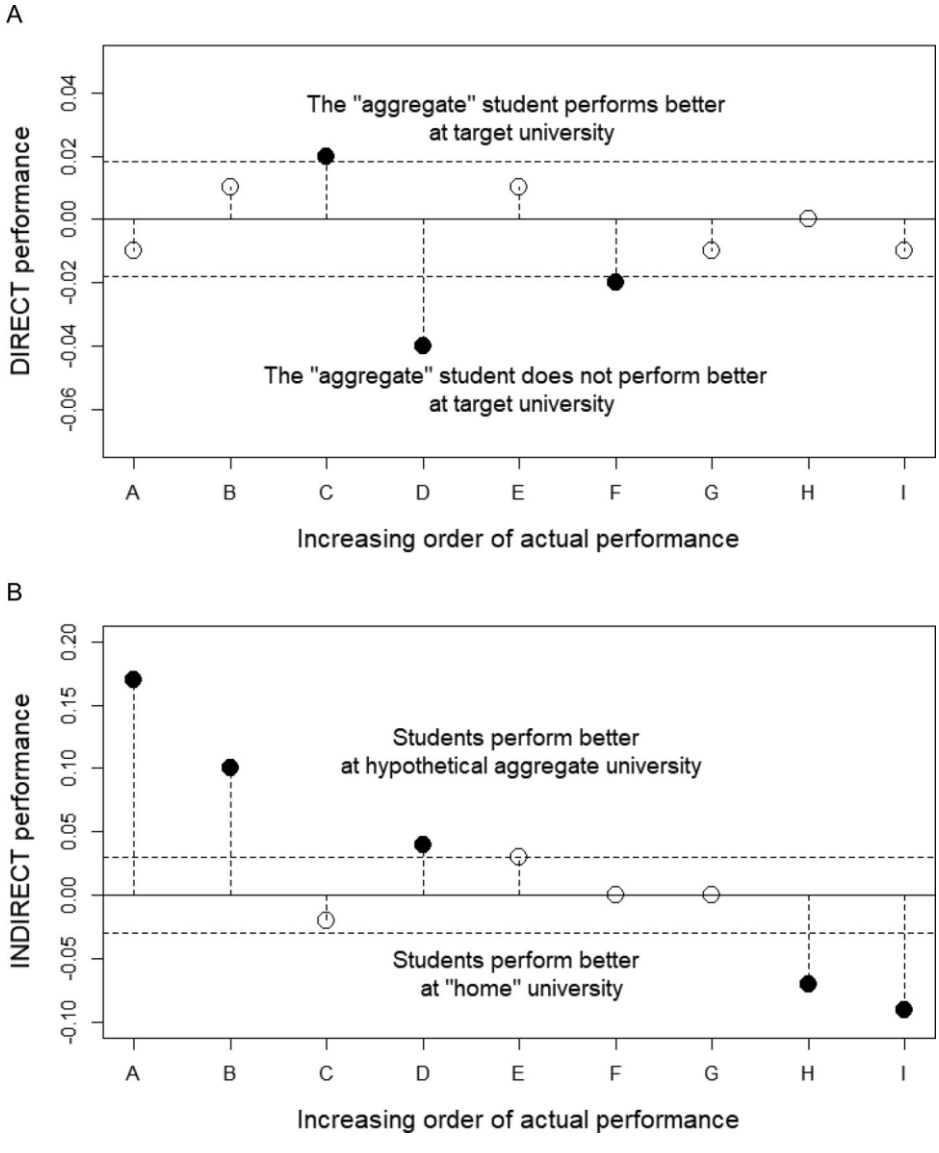

Figure 2. Results of benchmarking analysis for nine institutions. The *Y* axis measure of relative performance is the actual average achievement of students minus the expected benchmark achievement for each benchmarking method, respectively. (a) Shows results for direct standardization performance. (b) Shows results for indirect standardization performance. The solid lines indicate the mean benchmark performance across all institutions. The dashed line indicates 1 standard deviation of mean benchmark performance across all institutions.

benchmarks showed no evidence of a relationship between the benchmark and actual performance. In simple terms, outperforming an IS benchmark indicates that more students are successfully completing their studies at a particular institution than would be expected on average at all other universities combined (i.e. the hypothetical 'aggregate' university). For DS, the interpretation is suggested to be more nuanced. Thus, an institution with a graduation rate that is higher than their DS benchmark suggests that the wider, aggregate student population would be expected to be more successful at the benchmarked institution than the student population that actually studied there. Compared to benchmarking practices in the business sector, the use of benchmarking in HE is in its infancy. While there are established frameworks for benchmarking practices in business research (Rolstadas

2013), the use and interpretation of benchmarking in HE has been comparatively limited (Chinta, Kebritchi, and Elias 2016). The main benchmarking approach used in HE has been IS, and its merits have been strongly argued (Draper and Gittoes 2004; Asif and Raouf 2013). Our results suggest that novel, complementary information might be gleaned from simultaneously considering DS benchmarking results, while not adding significantly to the complexity of the underpinning analysis. There is an accompanying need for a framework for interpretation of both DS and IS results for different stakeholders. Here, we propose a simple, generalized approach for interpreting results of benchmarking in HE for both IS and DS benchmarks, enabling stakeholders to distinguish unique characteristics of each approach.

First, we provide a conceptual model of benchmarking outcomes (Figure 3), showing how all benchmarks are interpreted relative to actual performance. To interpret deviations between benchmark and actual performance, there must be explicit consideration of both the possible outcomes, and also the meaning of specific deviation from actual performance for each respective benchmark approach. The latter is particularly important if, as in our results, there is no correlation between competing benchmarking outcomes. The dashed line in Figure 3 represents a baseline of actual performance relative to benchmarking outcomes. Three possible outcomes of benchmarked performances are shown as below, above or about the same as actual performance. Secondly, we need to revisit the threshold for the decision as to whether the benchmarked institution differs sufficiently from the actual performance. A difference of 0.03 (3%) has been previously suggested as a criterion for significance in IS benchmarking (Draper and Gittoes 2004). However, the arbitrary choice of a 3% difference may not be appropriate for DS benchmarks or for any benchmark with low variation.

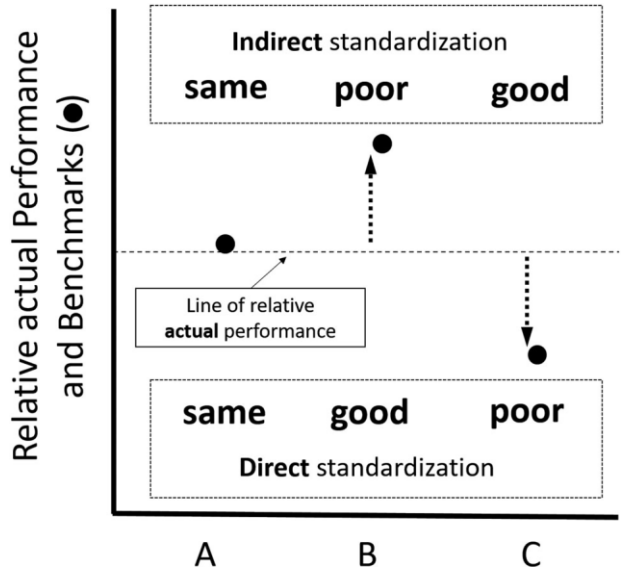

Figure 3. Scenarios for actual and benchmark performance outcomes for direct standardization (performance of the aggregate student demographic cohort at an individual institution) and indirect standardization (performance of the student cohort from an individual institution at the 'average national institution'). Three outcomes are possible: (1) Scenario A occurs when benchmark and actual performance are the same or negligibly different. The outcome in this case is the same (negligible effect represented by '0') for both direct and indirect benchmark methods. (2) Scenario B is when benchmark performance is significantly above that of actual performance. For direct benchmarking this is a positive outcome ('+'), where the average student would be expected to perform higher, relative to actual performance. For indirect benchmarking, this is a negative outcome ('–'), where students would be expected to perform better at the average national institution, compared to their own institution. (3) Scenario C is when benchmark performance is significantly lower than actual performance. For direct benchmarking, this is a negative outcome ('–'), where students from the average student cohort would be expected to perform less well at the target institution. For indirect benchmarking, this is a positive outcome ('+'), where students from the target institution would be expected to perform less well at the average national institution than at their own institution.

Here, we suggest a convenient measure is one standard deviation of the mean difference between actual and benchmark performance. When there is no significant difference from actual performance, this indicates no effect of the factor subdivisions on benchmark performance. This outcome is the same for both DS and IS. For DS benchmarks, the interpretation of no difference between the benchmark and the actual performance would be that the aggregate student population is expected to perform 'about the same' as the actual student performance at the benchmarked institution. For IS benchmarking, the student population from an individual institution would be expected to perform about the same as if they had attended the aggregate university.

We suggest that interpretation is different for the two benchmarking approaches, and as a consequence may be of different value to different stakeholders. For IS, when an institution outperforms its IS benchmark (top box in Figure 3), this can be interpreted as a 'good' outcome. Here, expected performance of the institution-specific cohort would be expected to fare less well elsewhere (on average). For the particular metric under scrutiny (in our case study, qualification rates), the demographic segments represented by the benchmark sample can be suggested to be 'well-suited' to the systems in place at that institution (such as the teaching and learning approaches). From the perspective of a student applying to this university, this can be interpreted as a positive expectation of outcome if choosing to attend the institution. However, this interpretation from the student perspective may be practically limited because the actual benchmark performance is only expected relative to the cohort of students that attended that university when the benchmark was calculated (Department for Education 2018). From a management or planning perspective, IS benchmarking could for example be used to explore and improve demographic segments that perform relatively less well locally compared to other institutions and inform a targeted management to improve performance within a particular student grouping. We also found that our IS benchmarks were significantly correlated with actual performance measures. This is because the more successful a particular institution is in metric performance (perhaps suggesting they are 'doing something right'), the more they outperformed their IS benchmark. However, correlation between IS benchmarks and actual performance suggests some redundancy in the information that the IS benchmark provides, as they largely mirror the pattern of absolute performances. Further enquiry is needed to confirm this pattern and understand the underpinning factors associated with actual performance and subsequent implications of the value of a sole reliance on IS benchmarking outputs that are strongly linked to actual performance.

For DS benchmarking (bottom box in Figure 3), the interpretation of deviation of benchmark scores from actual performance could be interpreted differently to benchmarks. Here, a benchmark score that is higher than actual performance could be interpreted as 'good', while a benchmark score lower than actual performance may be interpreted as 'poor'. This is because DS outcomes estimate the performance of the aggregate student population across all participating institutions were they to attend the benchmarked institution (Figure 1(a)). However, interpretation is perhaps more subtle than for IS. For example, in the scenario that institutions wanted to attract students in future from the wider student population, it could be argued that the expected performance of those aggregate demographic segments would be favourable compared to competing institutions. This interpretation does depend on the relative representation of those student demographic segments compared to the aggregate population, and this is based on the potential of reaching in future the wider student populations that would result in stronger performance. However, we suggest that this is an important consideration for university managers and practitioners that is only flagged by the use of DS benchmark approaches. Certain institutions may want to attract students in a widening participation context or perhaps international students who may not be relatively highly represented in a given student population. The outcomes could provide evidence that this would suit the institutional practices. However, if both the IS and (simplistic) DS benchmark outcomes were indicating a current situation of poor performance, then this is only an indicator that there is capacity for improvement through modifying the demographics of the resident student population.

As an example, we have applied this level of interpretation to our own regional results. When the outcomes of both DS and IS benchmarks are compared, several outcomes are apparent. First, in comparing outcomes for DS and IS benchmark outcomes (Figure 2(a,b), respectively), institution 'A' recorded a DS benchmark performance that was not different (less than 1 standard deviation) than actual performance, but IS benchmark performance was much higher than actual performance. While university A achieved the *status quo* for the DS benchmark, this is a poor outcome for the IS benchmark. Here, the student demographic would be expected to have performed better outside of the institution they actually attended. To stakeholders managing benchmark performance at this institution, this outcome could be used to improve outcomes (e.g. by managing improved performance). For potential applicants, it could inform relative expected outcomes at that institution, with the aforementioned caveat that the expectation is based on a past demographic cohort. Institution 'D' shows a very low DS benchmark (Figure 2(a); a negative outcome) and a slightly high IS benchmark (Figure 2(b); a negative outcome). Here, the interpretation of the DS benchmark is that the aggregate student population would be expected to perform less well at the target university than on average. This could indicate a 'good match' between programme management and their particular student population. However, performance would be expected to decrease were the demographic segments attending this institution to become more similar to those represented in the aggregate for the region (for example due to a change in applicant demographics or recruitment practice). We note that only institution D in our dataset showed significant deviation in benchmark performance for both IS and DS measures, and it suggests a poor outcome overall.

Conclusions

There is increasing international demand for data-led performance metrics in HE institutions. A promising approach in objective performance evaluation is benchmarking. Our findings suggest that both DS and IS benchmarks can add value to the information available for decision-making in HE. We suggest that the 'double-negative' benchmarked outcome will be of particular interest to wider stakeholders, such as funding bodies as these comparatively underperforming institutions appear to not serve student cohorts to a level seen in their sector. IS benchmarks are relatively straightforward to explain in terms of relative performance in the sector, whereas DS outcomes require more careful interpretation. We suggest that decision-making using both benchmarks should vary according to the type of institution and the types of students that they recruit and support. We have provided an interpretation framework for both benchmarks using an example of qualification rates in regional nursing and allied health degree programmes, supported by graphical explanations of the benchmarking approaches. We suggest that this example can be generalized across subject areas and international schema. While we have highlighted benchmarking across student demographic segments in our example here, we note that benchmarking segmentation can be used to highlight difference amongst other groupings, such as subject area (e.g. Department for Education 2018), or other regional or socio-economic factors that we do not explicitly consider here.

Benchmarking has value in accounting for the variation in institutions to make meaningful comparisons, when outcomes are interpreted with sufficient context (Rolstadas 2013; Hazelkorn 2015). Provisions of both IS and DS benchmarks should facilitate institutions to inform and enhance current practices. For IS outcomes at a particular institution, it is straightforward to interpret that if your students had attended a hypothetical 'aggregate' university and would have performed relatively better, then your institution is not performing well in the sector. However, for DS outcomes, different institutions may need to interpret the benchmarks in light of their ambitions for serving their students and the types of future students that may wish to recruit. Universities with a less diverse student population might interpret that an aggregate university cohort (composed of representative student group across the sector) would fare less well, as an indication that their academic provision is well-tailored to the needs of their own students. Universities with a narrow range of student typologies may interpret poor performance of an aggregate student at their own institution

as a success. This is because they specialize in teaching their own student typologies. However, universities with a high diversity of student typologies (e.g. social backgrounds), may want to interpret their DS metrics to enhance the likelihood of success of students represented at a national demographic level, with implications for good outcomes across different student socio-economic backgrounds and for broadening participation in HE.

The use of benchmarks in HE is increasing both in prevalence and importance (Agasisti and Bonomi 2014; Hazelkorn 2015). We suggest that the information provided by IS and DS outcomes can be complementary. The value to stakeholders of using both adjusted measures provides a stronger evidence base for decision-making, if tailored to the institutional typology and ambition. Benchmarking tools have been used widely in HE to monitor performance amongst institutions, for example in national ranking schemes or to inform management at the national level. These approaches seem better suited to IS benchmarking approaches that are relatively straightforward to interpret and compare. DS benchmarks contain different and relevant information that require context and interpretation from a particular institution's perspective. The outcomes of DS benchmarking could inform institutional ambitions to attract and teach students most suited to their educational systems. We therefore suggest that national-level reporting of both IS and DS benchmarking results can be useful to identify levels of relative performance of institutions, improve performance of specific student demographic segments and therefore be of value to national education policy, university managers, educational practitioners and students.

Acknowledgements

Benchmarks were created through consultation and agreement with the representatives of the Council of Deans from participating universities. The authors thank them and the representatives of Health Education North West, particularly Libby Sedgley, for consultations during the project development. Ethical approval for this study was obtained from the Manchester Metropolitan University, Faculty of Health, Psychology and Social Care Research Ethics Committee.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This project was funded by Health Education England.

References

Agasisti, T., and F. Bonomi. 2014. "Benchmarking Universities' Efficiency Indicators in the Presence of Internal Heterogeneity." *Studies in Higher Education* 39: 1237–55.

Asif, M., and A. Raouf. 2013. "Setting the Course for Quality Assurance in Higher Education." *Quality and Quantity* 47: 2009–24.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Bowden, R. 2000. "Fantasy Higher Education: University and College League Tables." *Quality in Higher Education* 6: 41–60.

Chinta, R., M. Kebritchi, and J. Elias. 2016. "A Conceptual Framework for Evaluating Higher Education Institutions." *International Journal of Educational Management* 30: 989–1002.

Department for Education. 2018. *Teaching Excellence and Student Outcomes Framework: Subject-Level*. London, UK: Crown Copyright.

Dill, D. D., and M. Soo. 2005. "Academic Quality, League Tables, and Public Policy: A Cross-National Analysis of University Ranking Systems." *Higher Education* 49: 495–533.

Draper, D., and M. Gittoes. 2004. "Statistical Analysis of Performance Indicators in UK Higher Education." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167: 449–74.

Fielding, A., P. J. Dunleavy, and A. M. Langan. 2010. "Interpreting Context to the UK's National Student (Satisfaction) Survey Data for Science Subjects." *Journal of Further and Higher Education* 34: 347–68.

Genuer, R., J-M. Poggi, and C. Tuleau. 2010. "Variable Selection Using Random Forests." *Pattern Recognition Letters* 31: 2225–2236.

Hall, M. A., and G. Holmes. 2003. “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining.” *IEEE Transactions on Knowledge and Data Engineering* 15: 1437−47.
Hapfelmeier, A., and K. Ulm. 2013. “A New Variable Selection Approach Using Random Forests.” *Computational Statistics & Data Analysis* 60: 50−69.
Harrison, N., and S. Hatt. 2010. ““Disadvantaged Learners”: Who Are We Targeting? Understanding the Targeting of Widening Participation Activity in the United Kingdom Using Geo-Demographic Data From Southwest England.” *Higher Education Quarterly* 64: 65−88.
Hazelkorn, E. 2007. “The Impact of League Tables and Ranking Systems on Higher Education Decision Making.” *Higher Education Management and Policy* 19: 1−24.
Hazelkorn, E. 2015. *Rankings and the Reshaping of Higher Education: The Battle for World-Class Universities*. 2nd ed. Basingstoke: Palgrave-Macmillan.
HEFCE. 1999. Performance Indicators in Higher Education in the UK.
HESA. 2011. International Benchmarking in UK Higher Education.
Jackson, N. 2001. “Benchmarking in UK HE: An Overview.” *Quality Assurance in Education* 9: 218−235.
Jackson, N., and H. Lund. 2000. *Benchmarking for Higher Education*. Milton Keynes, UK: Open University Press.
JISC. 2012. Benchmarking: Key tools to develop your understanding and use of benchmarking. Accessed 15 February 2018. https://www.jisc.ac.uk/full-guide/benchmarking
Kumar, S., and C. Chandra. 2001. “Enhancing the Effectiveness of Benchmarking in Manufacturing Organizations.” *Industrial Management & Data Systems* 101: 80−89.
Kyrö, P. 2003. “Revising the Concept and Forms of Benchmarking.” *Benchmarking: An International Journal* 10: 210−25.
Langan, A. M., W. E. Harris, N. Barrett, C. Hamshire, and C. Wibberley. 2016. “Benchmarking Factor Selection and Sensitivity: A Case Study with Nursing Courses.” *Studies in Higher Education* 2: 1−11.
Liaw, A., and M. Wiener. 2002. “Classification and Regression by RandomForest.” *R News* 2/3: 18−22.
Northcott, D., and S. Llewellyn. 2005. “Benchmarking in UK Health: A Gap Between Policy and Practice?” *Benchmarking: An International Journal* 12: 419−35.
Pitts, D. W. 2005. “Diversity, Representation, and Performance: Evidence about Race and Ethnicity in Public Organizations.” *Journal of Public Administration Research and Theory* 15: 615−31.
R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rolstadas, A. 2013. *Benchmarking — Theory and Practice*. New York: Springer.
Shober, A. F. 2013. “Debate: Benchmarking Inequality—Driving Education Progress in the USA.” *Public Money & Management* 33: 242−4.
Tam, M. 2001. “Measuring Quality and Performance in Higher Education.” *Quality in Higher Education* 7: 47−54.
Tee, K. F. 2016. “Suitability of Performance Indicators and Benchmarking Practices in UK Universities.” *Benchmarking: An International Journal* 23: 584−600.
Verikas A., A. Gelzinis, and M. Bacauskiene. 2011. “Mining Data with Random Forests: A Survey and Results of New Tests.” *Pattern Recognition* 44: 330−349.
Williams, R., and G. de Rassenfosse. 2016. “Pitfalls in Aggregating Performance Measures in Higher Education.” *Studies in Higher Education* 41: 51−62.
Zairi, M. 1998. *Benchmarking for Best Practice: Continuous Learning Through Sustainable Innovation*. London, UK: Routledge.